Studies in Second Language Learning and Teaching
Department of English Studies, Faculty of Pedagogy and Fine Arts, Adam Mickiewicz University, Kalisz
SSLLT 5 (1). 2015. 153-169
doi: 10.14746/ssllt.2015.5.1.8
http://www.ssllt.amu.edu.pl

The handling of missing binary data in language research

François Pichette
Université du Québec- Téluq, Canada
francois.pichette@teluq.ca

Sébastien Béland
Université de Montréal, Canada
sebastien.beland@umontreal.ca

Shahab Jolani
Universiteit Utrecht, the Netherlands
s.jolani@uu.nl

Justyna Leśniewska
Uniwersytet Jagielloński, Poland
justyna.lesniewska@uj.edu.pl

Abstract

Researchers are frequently confronted with unanswered questions or items on their questionnaires and tests, due to factors such as item difficulty, lack of testing time, or participant distraction. This paper first presents results from a poll confirming previous claims (Rietveld & van Hout, 2006; Schafer & Graham, 2002) that data replacement and deletion methods are common in research. Language researchers declared that when faced with missing answers of the yes/no type (that translate into zero or one in data tables), the three most common solutions they adopt are to exclude the participant's data from the analyses, to leave the square empty, or to fill in with zero, as for an incorrect answer. This study then examines the impact on Cronbach's α of five types of data insertion, using simulated and actual data with various numbers

of participants and missing percentages. Our analyses indicate that the three most common methods we identified among language researchers are the ones with the greatest impact on Cronbach's α coefficients; in other words, they are the least desirable solutions to the missing data problem. On the basis of our results, we make recommendations for language researchers concerning the best way to deal with missing data. Given that none of the most common simple methods works properly, we suggest that the missing data be replaced either by the item's mean or by the participants' overall mean to provide a better, more accurate image of the instrument's internal consistency.

Keywords: missing data, Cronbach's alpha, participant exclusion, second language testing

1. Introduction

Language research often involves the use of questionnaires that consist of open-ended as well as multiple-choice or yes-no questions. It is typical for researchers to collect incomplete data, since participants often leave a certain number of items unanswered. Several factors can account for such missing data: fatigue (e.g., last items of a long questionnaire), distraction (e.g., the back side of a copy left blank), item difficulty (items skipped or ignored), and so on.

There are ways to avoid missing data, a common one being the forcing of answers in computerized tests. However, it may not always be desirable, since we want and expect participants to answer to the best of their knowledge. Forcing answers may impact reliability; therefore, as much as we would not include reluctant participants who would perform the whole tasks negligently or carelessly, we may want to avoid forcing unwilling participants to do the same for parts of our tasks. Missing data may thus be a necessary evil.

How exactly researchers in language studies deal with this necessary evil, however, is unclear. There are no established procedures, and few recommendations exist that could provide guidance to researchers who need to address this problem, despite the fact that a large number of solutions to the problem of missing data have been developed by statisticians (which vary greatly in terms of their complexity).

The aim of this paper is twofold. Firstly, we try to obtain a picture of the current situation with respect to how missing data are handled by language researchers. Secondly, we try to establish which of the available procedures for handling missing data is the best, as well as the most practical and realistic solution for language researchers.

2. The handling of missing data in language research: The current situation

Several solutions are found in the literature for dealing with missing data (see, e.g., Schafer & Graham, 2002). The following are simple solutions used with items that require scoring (Peugh & Enders, 2004). The first two are deletion methods:

- eliminating from the analyses the participants whose data is missing;
- disregarding the missing data with no figure for replacement.

The following ones are replacement methods:

- replacing the missing data by zero;
- replacing by the participant's mean;
- replacing by the mean score for the item;
- replacing by the overall mean for the test, and so on.

Despite missing data being recognized as "a common problem" (Blom & Unsworth, 2010, p. 3), in a large number of applied linguistics studies the handling of missing data is not addressed, or no rationale is provided for the solution adopted by the researchers. We conducted an analysis of articles published in four journals related to language and linguistics over a period of five years (2007-2011). Each journal was in the middle of each of the four quartiles for impact on SCImago (2014),[1] that is, at or around the rankings of 50, 150, 250 and 350, with respective impact ratings of .68, .20, .11 and .10.[2]

The total number of articles in the four journals was 278, of which 221 (79.5%) had exclusively theoretical content and did not involve participants. Of the remaining 57 studies that involved participants, only two (3.5%) mentioned the exclusion of participants, one article in the first quartile journal and one in the third. In addition, of those 57 studies, which are overwhelmingly qualitative,[3] only two alluded to the existence of missing data, both in the first quartile

[1] The *language and linguistics* category was used, comprising around 400 journals. There was an unequal number of items across quartiles, which explains why the ranks of the four journals are not equidistant. In the case of an even number of journals in a quartile, the first of the two journals in the middle of the quartile was selected. We intentionally provide approximate figures for rankings to keep the journals anonymous.

[2] The third and fourth journal offer a potentially equal and low impact, but that selection method reflects the actual bulk of publications in the field. Selecting leading journals would have misrepresented the field and might have skewed our data toward more statistically inclined researchers and quantitatively sound analyses. A sampling search in bibliographic databases for studies employing particular types of datasets would have prevented us from obtaining a broad view of what is published in the field.

[3] More than 90% of studies in language learning are said to be of a qualitative nature (Lazaraton, 2000) although this figure is said to be declining (Sterling, Wolff, & Papi, 2012). Such a high percentage is also true for social studies in general, reflecting a sudden increase in qualitative research, which more than doubled in the 1990s and continues to predominate (Fallon, 2006; Shaver, 2001).

journal. One article mentioned that some data were inaudible but could be reconstructed from memory; the other article referred to data that was rejected because of the participants' poor age-related performance. Given that data collection does not always go exactly as planned, and that missing data is a common problem, as was just mentioned, the fact that only two studies out of 57 allude to missing data suggests that in most cases, researchers do not mention how they dealt with their missing data, or when they do so, no rationale is provided for the solution they adopted.

Four journals do not provide a complete picture of the situation, since each of them may have been heavily influenced by individual editorial policy or statistical reviewing (or lack thereof). While an analysis of a larger number of journals would doubtless provide a more accurate image of data handling in language research, the present analysis seems sufficient to claim that the existence and handling of missing data seems to be avoided by language researchers as a problematic topic.

There seem to be no recommendations in the literature which would make it clear to an average researcher in the field of language studies which of the aforementioned deletion and replacement methods to use, and under which circumstances. Journal articles do not provide an established standard which researchers could follow.

In addition, the avoidance of the topic in studies involving data collection from human participants could be attributed to lack of statistical training. As stated by Yang (2010, p. xii), the UK's Social Research Council "has long recognized the lack of statistics training among the British graduates," many of whom become language researchers. In a questionnaire recently completed by 380 language researchers (Schmidtke, Spino, & Lavolette, 2012), only about 30% of professors considered their training in statistics to be adequate, 80% of the respondents had never had a statistics class, and 12% do not compute statistics. These figures suggest that little has changed since Lazaraton, Riggenbach, and Ediger's (1987) study of 121 language researchers, which showed that about half of them had taken either zero or one class in statistics or in research methods.

In the course of this study, an invitation to fill out an online questionnaire was sent out to 810 people who conduct research related to languages, whose email addresses were compiled from recent language conference programs. The questionnaire was completed by 99 respondents. None of the respondents was excluded from our analyses. Of those respondents, about half were faculty members (n = 50) while the rest were graduate students. All 99 respondents had been presenters at scientific conferences. Our survey consisted of four multiple-choice questions, each with an option for addition, plus one open-ended question. Two identification questions were asked about the participants' academic status and research discipline.

Excluding participants is or has been a solution for about 87 percent of the respondents. The three main answers chosen as reasons for excluding participants are as follows:

- Their scores or answers were too different from the other participants' (29%).
- They did not answer all questions or items. I want or need each participant to answer all questions (12%).
- They had too many answers missing. I want or need a certain number of answers from each participant (14%).

Of the respondents who excluded participants due to missing answers or to their scores, about half (48%) indicate they have no preset criteria for exclusion, while the other half (52%) say they do. For exclusion based on scores, three respondents require a gap of 2 standard deviations (*SD*) from the mean, one requires 2.5 *SD*s, and three require 3 *SD*s. Another participant sets the cut-off point at 20% from the mean. For exclusion based on the number of missing items, the range is even wider: Three respondents exclude participants who reach 5% of missing answers, while another respondent sets the threshold at 50%. However, understandably, most reasons for excluding participants are of a demographic (inappropriate profile) or contextual nature (absence from some tests, dropping out of a longitudinal study, etc.). Demographic exclusion is related to unwanted data, while contextual exclusion leads to missing data.

Among our respondents, 31% claim they never have to perform calculations on matrices with missing data. For the respondents who work with data matrices, the solutions they have adopted are:

- leaving the square empty (39%);
- inserting that participant's mean for the rest of the items (12%);
- inserting zero (11%);
- inserting the mean score obtained by everyone for that item (10%);
- inserting the mean on the whole test (1%).

In studies with dichotomous data (i.e., when the possible answer is scored as either 0 or 1), some participants are of the opinion that a missing answer can be replaced by a 1 when the researcher has strong reasons to believe the respondent would have obtained this score had he or she answered the question or done the item that was overlooked.

The results of the questionnaire show clearly that there is a wide diversity in the way researchers deal with missing data. The results of the survey are also very interesting when contrasted with the findings from the investigation of published journal articles mentioned earlier: While 87% of our respondents admit to excluding participants, only 3.5% of researchers report doing so in published articles. This situation was touched upon by several of our respondents, which is exemplified by the following comment: "*My primary concern, at least at the moment, is lack of a*

sufficiently detailed description in many published articles of how the data was actually handled and analyzed." One likely explanation may reside in another comment: "*It feels that if you're honest about how you exclude participants, your paper is less likely to be accepted because it raises too many questions*."

Another obvious gap appears between the high number (52%) of our respondents who collect data, and the 20% of published studies that contain collected data. The reason for this gap, however, is most likely due to the fact that researchers who collect data and who show an interest in statistics may be overrepresented among our respondents. Many people who were contacted declined to answer our questionnaire, arguing that they do qualitative research and/or do not collect data for their research. This was made known to us by personal emails sent in response to the request to fill in the questionnaire.

The treatment of missing data in language studies is thus a serious issue, to which little attention has been drawn. This ongoing situation is surprising in light of the amount of research on the treatment of missing data in other fields of study such as psychology or sociology, and outside the social research sphere, mostly in natural sciences (e.g., see Allison, 2001). Table 1 provides examples of such studies for various fields of social sciences.

Table 1 Studies on missing data in social research

Fields of social science	Some studies
Economics	Florez-Lopez (2010); Graham (2011); Harvey & Pierse (1984); Horowitz and Manski (1998); Nicoletti, Peracchi, and Foliano (2001); Philipson (2001)
Political science	King, Honaker, Joseph, and Scheve (2001); Tufis (2008)
Psychology and education	Finch (2008); Peugh and Enders (2004); Robitzsch and Rupp (2009); Schafer and Graham (2002); Zhang and Walker (2008)
Sociology	Allison (1987); Schrapler (2004); Winship and Mare (1989)

3. In search of a practical solution to the problem of missing data

To address the problem of missing data, statisticians have provided dozens of techniques, which range from simple solutions like the deletion or simple replacement methods mentioned above to more advanced statistical procedures such as maximum likelihood or multiple imputation methods (see, e.g., Little & Rubin, 2002 or Schafer & Graham, 2002). Such techniques may have their advantages; however, since they require advanced statistical knowledge, it would be unrealistic to expect that they can be widely adopted by language researchers. Rather, most researchers would benefit more from information about which of the simple, easily accessible options they are familiar with is the best one to use when dealing with missing data. Our aim in this paper is to try to provide an answer to this question.

Researchers traditionally discard missing data because this method is very easy to implement and almost all statistical packages offer it as an option. The most important consequence of removing missing values, however, is reduction in sample size which results in a loss of statistical power. Nevertheless, deletion methods have enjoyed widespread use in all disciplines (Marsh, 1998; Peugh & Enders, 2004).

It has to be remembered that value replacement adds data points that are not real, which always results in some kind of data fabrication. Ideally, all incomplete data sets should be eliminated. Realistically, if researchers did so, whole studies would be jeopardized in cases where many of the participants have at least one missing answer. This solution (listwise deletion) also results in the loss of useful information. As was shown by our poll, language researchers will eliminate data sets (i.e., all data for a participant) only beyond a certain threshold of missing answers. Below that threshold, the missing values are simply dealt with instead of being discarded.

Instead of excluding data from participants who did not answer all the questions or items on a test, especially when very few answers are missing, many language researchers opt for filling in each missing value with a replacement value. As opposed to deletion methods, this solution is appealing since the incomplete data sets (i.e., matrices in linguistics) are converted to complete data sets. Convenience therefore is the major benefit of using such methods, and any statistical analyses can be applied to the completed data sets. Despite these apparent advantages, any replacement method might have potential drawbacks by producing biased results because of uncertainty about which value to insert. This also causes researchers to underestimate variance (Little & Rubin, 2002, p. 61).

Even though there has been considerable research on missing data by statisticians and psychometricians, comparisons between imputation (i.e., fill-in) methods are normally made using simulated data with very large data sets, and involving complex methods. Very few comparisons, if any, have been made between the various simple imputation methods actually used in research, and for a realistic number of participants and items.

In order to find out which of the commonly used, realistic methods of dealing with missing data in language research is the best one, we formulated the following more specific research goals:

1. To compare the impact on Cronbach's α of seven common fill-in methods for normal-size matrices, using simulated data.
2. To compare real research data to simulated data.

The next sections will explain every step of our procedure.

4. Method

4.1. Procedure

Our procedure was based on the collection and analysis of dichotomous data, represented numerically in matrices as either 1s or 0s. This is the type of data collected in language research when questionnaires contain true/false or yes/no questions, for example.

We examined the impact on Cronbach's α coefficients of the following five single fill-in methods: replacement of the missing data (a) by 0, (b) by 1, (c) by the participant's mean on all items, (d) by the item's mean for all participants, and (e) by the overall mean on the test for all participants.

Cronbach's α coefficient (Cronbach, 1951) was chosen in this study because it is by far the most commonly used measure for assessing the internal consistency of tests and questionnaires in social language research. For example, Sijtsma (2009) writes that "probably no other statistic has been reported more often as a quality indicator of test scores than Cronbach's (1951) α coefficient" (p. 107). Furthermore, Peterson (1994) states that "not only is coefficient alpha the most widely used estimator of reliability, but also it has been the subject of considerable methodological and analytical attention" (p. 382). It must be noted here that the current widespread use of Cronbach's α does not mean that the measure attracts unanimous approval from statistics experts; indeed, Cronbach's α has lately been the subject of criticism (see Sijtsma, 2009).

4.2. Instruments

4.2.1. Simulated data sets

We conducted a comprehensive simulation study to compare different single fill-in methods. To reach our two research goals, we created data sets for 20, 50, 250 and 500 participants and 20, 40 and 60 items. These combinations give matrices with different sizes, which are considered small, medium and large by current standards in language research. In order to reflect the range of α coefficients found in research papers, the binary matrices were generated such that the internal association among the items (Cronbach's α) varies from 0.4 to 0.9. Even though higher stakes assessments typically involve alphas upwards of .90, the range we selected is deemed to reflect levels that are encountered in actual research uses. For our study, we also chose percentages of missing answers of 5%, 10% and 20%. For each combination (item X participant X percentage), 1,000 matrices were generated, for a total of 12,000 matrices. The test matrices were generated using R.

For the sake of ecological validity, the simulated number of participants corresponded to the figures normally found in actual language research and was much lower than the thousands we usually find in papers published by statisticians. For example, in our review of published language papers mentioned earlier, 51 of the 57 studies that involved participants actually mentioned the number of participants. Of those 51 studies, more than half (that is, 26 out of 51) had fewer than 20 participants, and the average number of participants is 36.5, with a range of 1 to 217.

4.2.2. Matrices of test scores

To reach our second research goal, we retrieved two matrices with dichotomous data that were used in past experiments and that we consider representative of typical data collected in language research or testing.

The first matrix (Matrix A) contains data that was collected as part of a research project on reading ability. The test consisted of 64 yes-no questions and it was completed by 171 participants. For more information on the psychometric details of the test, see Pichette, Béland, Lafontaine, and de Serres (2014).

The second matrix (Matrix B) is a test that assesses the level of English as a second language of 1,709 students entering college (Quebec, Canada). This test contains 85 items where the students have either a right or a wrong answer. Shohamy, Donitsa-Schmidt and Ferman (1996) defined high-stakes tests as those which have important consequences and in which decisions about "admission, promotion, placement, or graduation are directly dependent on test scores" (p. 300). The test that provided data for Matrix B is an example of high stakes assessment because of its impact on students' academic path; hence its especially high Cronbach's α of .96. See Laurier, Froio, Paero, and Fournier (1998) for more information.

By pure coincidence, both matrices contained 0.4 percent of missing data. Random deletion was performed on those real data to reach the above mentioned percentages of missing answers (5, 10 and 20) to allow comparisons with our simulated data.

5. Results

In this section the results are presented in light of the two research goals stated earlier.

5.1. First research goal

This goal was to compare the impact on Cronbach's α of seven common methods of dealing with missing data for plausible numbers of participants, using

simulated data. The α coefficients yielded by each method are displayed in Tables 2 to 4. These tables represent average Cronbach's α coefficients computed over 1,000 iterations. In each table, "true α" represents the Cronbach's α coefficient calculated with a complete matrix, that is, without missing data. This coefficient is followed by the coefficients yielded by each of the five simple replacement methods that we selected, which in turn are followed by the two deletion methods described earlier.

Table 2 Simulated data, 20 items

				5%				10%				20%
Missingness	*N* = 20	*N* = 50	*N* = 250	*N* = 500	*N* = 20	*N* = 50	*N* = 250	*N* = 500	*N* = 20	*N* = 50	*N* = 250	*N* = 500
(True α)	.45	.48	.49	.50	.42	.47	.49	.49	.43	.47	.49	.49
Zero (0)	.50	.53	.54	.55	.54	.57	.58	.58	.60	.62	.63	.64
One (1)	.50	.53	.54	.55	.53	.57	.58	.58	.59	.62	.63	.64
Participant's mean	.48	.50	.52	.52	.48	.53	.55	.55	.54	.57	.59	.59
Item's mean	.44	.47	.49	.49	.41	.46	.49	.49	.40	.45	.47	.48
Overall mean	.44	.47	.49	.49	.41	.46	.49	.49	.40	.45	.47	.48
Listwise (CC)	.50	.53	.54	.55	.53	.56	.58	.58	.59	.62	.63	.63
Pairwise (AC)	.50	.53	.54	.55	.53	.56	.58	.58	.59	.62	.63	.63

Table 3 Simulated data, 40 items

				5%				10%				20%
Missingness	*N* = 20	*N* = 50	*N* = 250	*N* = 500	*N* = 20	*N* = 50	*N* = 250	*N* = 500	*N* = 20	*N* = 50	*N* = 250	*N* = 500
(True α)	.66	.68	.69	.69	.65	.68	.69	.69	.65	.68	.69	.69
Zero (0)	.70	.72	.73	.73	.72	.75	.75	.76	.77	.78	.79	.79
One (1)	.70	.72	.73	.73	.72	.75	.75	.76	.77	.78	.79	.79
Participant's mean	.67	.69	.70	.70	.68	.71	.71	.72	.71	.73	.74	.74
Item's mean	.65	.67	.69	.69	.64	.67	.68	.69	.64	.67	.68	.68
Overall mean	.65	.67	.69	.69	.64	.67	.68	.69	.64	.67	.68	.68
Listwise (CC)	.70	.72	.73	.73	.72	.75	.75	.76	.77	.78	.79	.79
Pairwise (AC)	.70	.72	.73	.73	.72	.75	.75	.76	.77	.78	.79	.79

Table 4 Simulated data, 60 items

				5%				10%				20%
Missingness	*N* = 20	*N* = 50	*N* = 250	*N* = 500	*N* = 20	*N* = 50	*N* = 250	*N* = 500	*N* = 20	*N* = 50	*N* = 250	*N* = 500
(True α)	.76	.77	.78	.78	.75	.77	.78	.78	.75	.76	.78	.78
Zero (0)	.79	.80	.81	.81	.81	.82	.83	.83	.84	.85	.85	.85
One (1)	.79	.80	.81	.81	.81	.82	.83	.83	.84	.85	.85	.85
Participant's mean	.76	.78	.79	.79	.77	.79	.79	.79	.78	.80	.81	.81
Item's mean	.75	.77	.77	.78	.74	.76	.77	.77	.73	.75	.77	.77
Overall mean	.75	.77	.77	.78	.74	.76	.77	.77	.73	.75	.77	.77
Listwise (CC)	.79	.80	.81	.81	.81	.82	.83	.83	.84	.85	.85	.85
Pairwise (AC)	.79	.80	.81	.81	.81	.82	.83	.83	.84	.85	.85	.85

It has to be noted that even though in most cases the two types of deletion and replacement by zero yield apparently identical Cronbach's α coefficients, this is due to the rounding up at the second decimal. In most cases, if not all, differences lie in the third or fourth decimal.

The following observations can be made from the above tables:

1. In a majority of cases (196 out of 252, or 78%), any bias induced by replacement or deletion leads to higher Cronbach's α coefficients, sometimes by as much as .17. The only methods that yield lower coefficients are the replacement by the item's mean and by the overall mean, but this decrease is always negligible, being -.01 every time it occurs.
2. For a given number of items, the bias tends to increase proportionally with the percentage of missing data.
3. For a given percentage of missing data, in absolute numbers, the bias tends to be similar no matter the sample size.
4. Bias tends to decrease when the number of items increases.

The methods that have the least impact on Cronbach's α coefficient are the replacement either by the item's mean or by the participants' overall mean on the test. Replacement by the participant's mean on the test induces notable bias on Cronbach's α coefficient. Replacing the missing data by zero or by one, or adopting either type of deletion are the methods that yield the strongest bias. Given the fourth observation above, these differences of impact between methods are more apparent when looking at matrices based on the smallest number of items (20 items).

5.2. Second research goal

The results from our two matrices of real data are presented in Tables 5 and 6. The true alpha coefficients in the absence of data are provided in table titles. The percentage of missing value of 0.4 was the real one, as mentioned earlier, while the other percentages were generated at random.

Table 5 Matrix A (real data), 64 items, *N* = 171, true α = .82

Missingness	0.4%	5%	10%	20%
Zero (0)	.83	.92	.92	.92
One (1)	.80	.84	.86	.88
Participant's mean	.81	.84	.84	.86
Item's mean	.81	.82	.83	.82
Overall mean	.81	.82	.82	.82
Listwise (CC)	.83	.94	.95	.96
Pairwise (AC)	.84	.94	.94	.95

Table 6 Matrix B (real data), 85 items, *N* = 1709; true α = .96

Missingness	0.4%	5%	10%	20%
Zero (0)	.96	.96	.96	.95
One (1)	.96	.96	.96	.96
Participant's mean	.96	.96	.96	.97
Item's mean	.96	.96	.96	.95
Overall mean	.96	.96	.96	.95
Listwise (CC)	.96	.97	.97	.97
Pairwise (AC)	.96	.97	.97	.96

Both matrices here have a number of items higher than the matrices with simulated data. The result of Matrix A in Table 5, however, is to some degree comparable to the simulation results presented earlier in Table 4, where the number of items and participants were 60 and 250, respectively. It is clear from Table 5 that for missing percentages of 5% and above, except for replacing by the item's mean and overall mean, the other methods, particularly deletion methods, overestimate Cronbach's α. Given our earlier observation to the effect that bias tends to decrease when the number of items increases, it is not surprising to see little or no bias in Table 6; no difference of impact between methods can be observed in the second real data (Matrix B, Table 6) because the number of items is very high (i.e., 85).

Table 7 Matrix A2 (simulated data), 64 items, *N* = 171, true α = .82

Missingness	5%	10%	20%
Zero (0)	.84	.86	.87
One (1)	.84	.86	.88
Participant's mean	.83	.84	.85
Item's mean	.82	.82	.81
Overall mean	.82	.82	.81
Listwise (CC)	.84	.86	.87
Pairwise (AC)	.84	.86	.87

Table 8 Matrix B2 (simulated data), 85 items, *N* = 1709, true α = .96

Missingness	5%	10%	20%
Zero (0)	.96	.96	.96
One (1)	.96	.96	.96
Participant's mean	.96	.96	.97
Item's mean	.96	.96	.96
Overall mean	.96	.96	.96
Listwise (CC)	.96	.96	.96
Pairwise (AC)	.96	.96	.96

For additional confirmation that the relative impact of replacement methods corresponds to simulation, we have simulated data matrices with the same

Cronbach's alpha and the same number of items as for our two real data sets (see Tables 7 and 8). The fact that these tables show similar data patterns confirm the conclusions we have reached regarding the relative merits of the replacement methods we are comparing.

6. Discussion

The most striking observation made in this study is that the three most common ways for language researchers to deal with missing data—replacing by zero, deleting the data or excluding the participant—happen to be the three methods that have the greatest impact on Cronbach's α coefficients.

Contrary to what our first research goal implied, it is not the matrix size per se that determines the value of various deletion or replacement methods, since the number of participants has little effect on that value, but rather the number of items on the test. For two matrices of the same size, for example one with 30 participants X 60 items and another with 60 participants X 30 items, the latter is more prone to show bias in Cronbach's α coefficients due to its lower number of items.

In light of the observations above, it can safely be assumed that because many instruments designed by language researchers for collecting data have a low number of items, prudence is recommended in dealing with missing data, especially if the percentage of missing data is at 5% or above.

7. Conclusion

The goal of this paper is to inform language research colleagues about the impact that the methods they commonly use for handling missing data has on the index they most commonly use to report on their instrument's internal consistency. Since the most common methods used by researchers for dealing with missing data tend to inflate Cronbach's α coefficients, it is probable that most instruments for which such coefficients are reported in published papers are less, sometimes far less, reliable than what the authors and their readers are led to believe.

For language researchers who gather numerical data but who do not have extensive training in statistics and are not provided with help from professional statisticians, our results suggests that missing data be replaced either by the item's mean or by the participants' overall mean in data matrices. This solution will provide readers with a better, more accurate image of the instrument's internal consistency. Another solution that was brought to our attention after the experiment and that would be worth investigating in future research is the replacement of the

missing data by the mean of the column plus the mean of the row, divided by two, as suggested by Winer (1971).

Readers interested in the several recent more complex approaches to dealing with missing data of various kinds will find a useful introduction in the article by Barladi and Enders (2010).

The study reported here is not without limitations. Firstly, we focus on the popular and widely used Cronbach's α, but it needs to be noted that, while this coefficient is the dominant reliability measure found in language research publications, other measures also exist.

Secondly, our investigation is limited in scope to dichotomous data. Among the more quantitative research carried out in linguistics and language research, a substantial body of research involves nondichotomous numerical data, typically in the form of Likert-scale scores out of 5 or 7. It would be interesting to investigate the impact of deletion and various replacement methods with such data.

Thirdly, the conclusions here stand for research data that are missing completely at random, referred to as MCAR in the literature. Although the causes for missing answers on our research tests are not always easy to determine, there are situations where data might also be missing at random (MAR) or not at random (MNAR). Future analyses for those other types of missingness assumptions should help us determine whether our conclusions apply to various kinds of missing research data.

References

Allison, P. D. (1987). Estimation of linear models with incomplete data. *Sociological Methodology*, *17*, 71-103.

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Barladi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*, 5-37.

Blom, E., & Unsworth, S. (Eds.) (2010). *Experimental methods in language acquisition research*. Amsterdam: Benjamins.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Curtis, D. A., & Harwell, M. (1996). *Training graduate students in educational statistics*. Paper presented at the annual meeting of the American Educational Research Association, New York, USA.

Fallon, D. (2006). The buffalo upon the chimneypiece: The value of evidence. *Journal of Teacher Education, 57*(2), 139-154.

Finch, W. H. (2008). Estimation of item response theory parameters values in the presence of missing data. *Journal of Educational Measurement*, *45*, 225-246.

Florez-Lopez, R. (2010). Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data. *The Journal of the Operational Research Society*, *61*, 486-501.

Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica*, *79*, 437-452.

Harvey, A. C. & Pierse, R. G. (1984). Estimating missing observations in economic time series. *Journal of the American Statistical Association, 79*, 125-131.

Horowitz, J. L., & Manski, C. F. (1998). Censoring of outcomes and regressors due to survey non-response: Identification and estimation using weights and imputation, *Journal of Econometrics, 84*, 37-58.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review, 95*, 49-69.

Laurier, M. D., Froio, L., Paero, C., & Fournier, M. (1998). *L'élaboration d'un test provincial pour le classement des étudiants en anglais langue seconde au collégial*. Québec, Canada: Direction générale de l'enseignement collégial, ministère de l'Éducation du Québec.

Lazaraton, A. (2000). Current trends in research methodology and statistics in applied linguistics. *TESOL Quarter*ly, *34*, 175–181.

Lazaraton, A., Riggenbach, H., & Ediger, A. (1987). Forming a discipline: Applied linguists' literacy in research methodology and statistics. *TESOL Quarterly, 21*, 263-277.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.

Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling, 5*, 22-36.

Nicoletti, C., Peracchi, F., & Foliano, F. (2011). Estimating income poverty in the presence of missing data and measurement error. *Journal of Business & Economic Statistics*, *29*, 61-72.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research, 21*, 381-391.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74,* 525-556.

Philipson, T. (2001). Data markets, missing data, and incentive pay. *Econometrica, 69*, 1099-1111.

Pichette, F., Béland, S., Lafontaine, M., & de Serres, L. (2014). Measuring L2 reading comprehension ability: The Sentence Verification Technique. *The Quantitative Methods in Psychology, 10*(2), 95-106.

Rietveld, A. C. M., & van Hout, R. (2006). *Statistics for language research: Analysis of variance*. Berlin: Gruyter Mouton.

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement*, *69*, 18-34.

Schafer, J. L., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.

Schmidtke, J., Spino, L. A., & Lavolette, B. (2012, October). *How statistically literate are we? Examining SLA professors' and graduate students' statistical knowledge and training*. Paper presented at the 31st Second Language Research Forum, Pittsburgh, USA.

Schrapler, J.-P. (2004). Respondent behavior in panel studies: A case study for income-nonresponse by means of the German Socio-Economic Panel (SOEP). *Sociological Methods & Research, 33*, 118-156.

SCImago. (2014, October 27). *SJR — SCImago journal & country rank*. Retrieved from http://www.scimagojr.com

Shaver, J. P. (2001). The future of research in social studies – For what purpose? In W. B. Stanley (Ed.), *Critical issues in social studies for the 21st century* (pp. 231-252). Greenwich, CT: IAP.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*(3), 298-317.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120.

Sterling, S., Wolff, D., & Papi, M. (2012, October). *Students' and professors' views of statistics in SLA – A call for a change?* Paper presented at the 31st Second Language Research Forum, Pittsburgh, USA.

Tufis, C. D. (2008). Multiple imputation as a solution to the missing data problem in social science. *Metode de cercetare, 1-2*, 199-212.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Winship, C., & Mare, R. D. (1989). Loglinear models with missing data: A latent class approach. *Sociological Methodology*, *19*, 331-367.

Yang, K. (2010). *Making sense of statistical methods in social research*. London: Sage.

Zhang, B., & Walker, C. M. (2008). Impact of missing data on person model fit and person trait estimation. *Applied Psychological Measurement, 32,* 466-479.